Statistics Coursework 1 (SA)

Part A

Question 1

A) Write R code to create the boxplot below of the expression levels of the "Cyclin A1 mRNA" from the golub dataset.

```
# extracting required data from golub
library(hopach); data(golub)
```

```
## Loading required package: cluster
```

```
## Loading required package: Biobase
```

```
## Loading required package: BiocGenerics
```

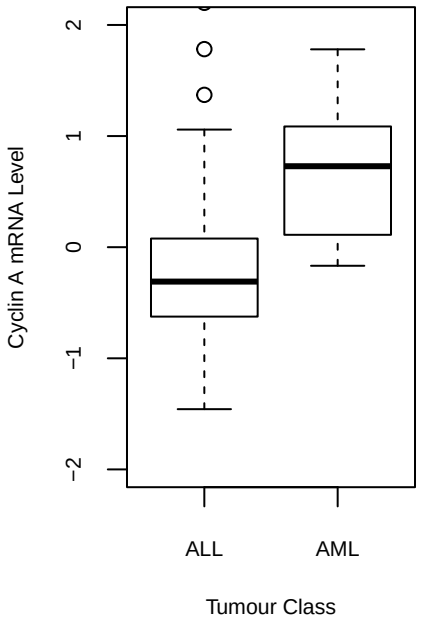
```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```r
tumour_class <- factor(golub.cl, levels=0:1, labels=c("ALL", "AML"))
golub_ALL <- golub[1553, tumour_class=="ALL"]
golub_AML <- c(golub[1553, tumour_class=="AML"], c(rep(NA, 16)))
golub_df <- data.frame(ALL = golub_ALL, AML = golub_AML)

# boxplot
par(ps=8, family="sans", mgp=c(2.5,1,0), pin=c(1.5,2.5))
box <- boxplot(golub_df, xlab="Tumour Class", ylab="Cyclin A mRNA Level",
               ylim=c(-2,2), col="white")
```



B) Test the hypothesis that the median expression levels of this gene differ between ALL and AML patients at a significance level =0.05.

Null hypothesis: The median expression level for AML and ALL patients is the same.

```r
# extracting the required data
tumour_class <- factor(golub.cl, levels=0:1, labels=c("ALL", "AML"))
golub_ALL <- golub[1553, tumour_class=="ALL"]
golub_AML <- golub[1553, tumour_class=="AML"]

# testing for difference of medians
wilcox.test(golub_ALL, golub_AML, alternative="two.sided", conf.level=0.95)
```

```
##
##  Wilcoxon rank sum exact test
##
```

```
## data:  golub_ALL and golub_AML
## W = 59, p-value = 0.003112
## alternative hypothesis: true location shift is not equal to 0
```

p-value (0.003112) is less than the significance level (0.05), therefore reject the null hypothesis - the median expression level of this gene differs between AML and ALL patients.

Question 2

A) Organize the data into a contingency table. Include the table in your answer.

236 overall, 82 with variant. 236-82 = 154 without variant in total.
Mutation, Survived: 42
Mutation, Died: 82 (total with variant)-42 = 40
No mutation, Died: 87 (total deceased)-40 = 47
No mutation, Survived: 154 (total without variant)-47 = 107

```
Survived <- c(42, 107)
Died <- c(40, 47)
contingency <- data.frame(Survived, Died, row.names=c("Mutation", "No Mutation"))
knitr::kable(contingency,
 caption = "Contingency table: Patient survival with and without gene X mutation")
```

Table 1: Contingency table: Patient survival with and without gene X mutation

	Survived	Died
Mutation	42	40
No Mutation	107	47

B) Use chisq.test() to determine whether survival is or is not dependent on the mutation of gene X. What is your conclusion?

Null hypothesis: Mutation status of gene X has no effect on the patient's survival outcome.

```
data <- matrix(c(42,40,107,47),2,byrow=TRUE)
chisq.test(data)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data
## X-squared = 6.9019, df = 1, p-value = 0.008611
```

```
qchisq(0.95,1)
```

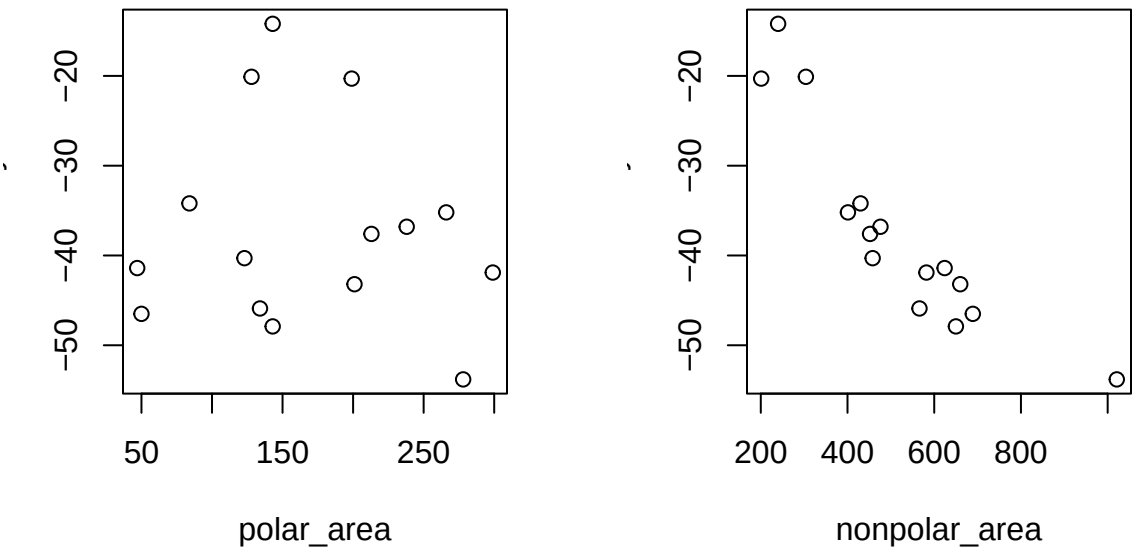
```
## [1] 3.841459
```

X-squared value (6.9019) is higher than the critical value of the chi squared distribution with 1 degree of freedom (3.84 - used because three out of the four marginal totals are used to construct expected values), therefore reject the null hypothesis: the mutation does have an effect on survival.

Question 3

A) Use correlation-based tests to examine the possible relationship between nonpolar area and affinity and between polar area and affinity. What can you conclude from these data?

```
nonpolar_area <- c(1021, 401, 582, 650, 201, 430, 689, 660, 624, 452, 458, 240,
                   304, 476, 566)
polar_area <- c(278, 266, 299, 143, 199, 84, 50, 201, 47, 213, 123, 143, 128,
                238, 134)
affinity <- c( -53.8, -35.2, -41.9, -47.9, -20.3, -34.2, -46.5, -43.2, -41.4,
               -37.6, -40.3, -14.2, -20.1, -36.8, -45.9)

# overview (quick plot)
par(mfrow=c(1,2), pin=c(2,2))
plot(polar_area, affinity)
plot(nonpolar_area, affinity)
```



```r
# relationship test - polar area vs affinity
cor.test(polar_area, affinity, alternative="two.sided", method="spearman")
```

```
## Warning in cor.test.default(polar_area, affinity, alternative = "two.sided", :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  polar_area and affinity
## S = 609.04, p-value = 0.7563
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.08757823
```

p-value (0.7563) is higher than the significance level of 0.05, therefore correlation coefficient rho is not significantly different from 0, therefore accept the null hypothesis: there is no correlation between polar area and binding affinity.

```r
# relationship test - non-polar area vs affinity
cor.test(nonpolar_area, affinity, alternative="two.sided", method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  nonpolar_area and affinity
## S = 1084, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.9357143
```

p-value is much lower than the significance level of 0.05, and the correlation coefficient is very close to 1 (-0.936) therefore reject the null hypothesis: there is a significant, strong negative correlation between non-polar area and binding affinity.

B) What were your reasons for selecting the particular statistical test or tests?

- The Pearson's correlation test assumes data is bivariate-normal and is strongly affected by outliers.
- Due to uncertainty about bivariate-normality, Pearsons may not be appropriate. Spearman's rank (non-parametric) is more appropriate, as bivariate-normal data is not assumed.
- From a basic plot, it is also possible that the non-polar data vs affinity data has an outlier point, and Spearman's rank is more robust to outliers than Pearson's correlation, therefore in case this is a statistically significant outlier, Spearman's rank was chosen.

Part B

Question 8

A) Investigate the relationships between individual enzyme concentrations and the metabolite level. On the basis of these relationships alone, for which of the enzymes is there evidence to support the idea that they are involved in controlling the level of the metabolite?

```
enzymeA <- c(0.114, 0.510, 0.723, 1.276, 1.93, 2.151,2.239, 2.732, 2.759, 3.015,
             3.617, 3.951, 6.281, 6.315, 6.693, 6.965, 7.057, 8.162, 8.216, 8.410)
enzymeB <- c(0.240, 4.011, 11.915, 12.393, 14.019, 15.047,19.418, 22.978, 23.483,
             29.311, 29.475, 31.647, 32.529, 32.827, 40.212, 44.497, 44.525, 47.422,
             49.717, 49.824)
enzymeC <- c(1.872, 3.177, 3.802, 4.028, 5.050, 5.227, 5.962, 8.211, 9.038, 9.392,
             9.738, 10.981, 12.145, 15.381, 16.897, 17.639, 17.744, 18.373, 18.692,
             19.350)
metaboliteB <- c(28.6, 35.8, 65.7, 135.2, 183.7, 229.2, 221.7, 256.9, 274.8, 312.0,
                 415.0, 490.4, 1014.2, 1085.9, 1222.2, 1308.6, 1351.1, 1680.8,
                 1722.0, 1793.5)
```

```
# enzymeA vs metaboliteB
enzymeA_test <- c(cor.test(enzymeA, metaboliteB, method="spearman")[[3]],
                  cor.test(enzymeA, metaboliteB, method="spearman")[[4]],
                  cor.test(enzymeA, metaboliteB, method="pearson")[[3]],
                  cor.test(enzymeA, metaboliteB, method="pearson")[[4]])

# enzymeB vs metaboliteB
enzymeB_test <- c(cor.test(enzymeB, metaboliteB, method="spearman")[[3]],
                  cor.test(enzymeB, metaboliteB, method="spearman")[[4]],
                  cor.test(enzymeB, metaboliteB, method="pearson")[[3]],
                  cor.test(enzymeB, metaboliteB, method="pearson")[[4]])

# enzymeC vs metaboliteB
enzymeC_test <- c(cor.test(enzymeC, metaboliteB, method="spearman")[[3]],
                  cor.test(enzymeC, metaboliteB, method="spearman")[[4]],
                  cor.test(enzymeC, metaboliteB, method="pearson")[[3]],
                  cor.test(enzymeC, metaboliteB, method="pearson")[[4]])

relationship_tests <- data.frame(enzymeA_test, enzymeB_test, enzymeC_test,
                                 row.names=c("spearman_p", "spearman_rho",
                                             "pearson_p", "pearson_cor"))
knitr::kable(relationship_tests,
 caption = "Relationship test outputs between metaboliteB and enzymes A, B, and C")
```

Table 2: Relationship test outputs between metaboliteB and enzymes A, B, and C

	enzymeA_test	enzymeB_test	enzymeC_test
spearman_p	0.0000060	0.0000060	0.0000060
spearman_rho	0.9984962	0.9984962	0.9984962
pearson_p	0.0000000	0.0000000	0.0000000

	enzymeA_test	enzymeB_test	enzymeC_test
pearson_cor	0.9813160	0.9292840	0.9633902

Using Spearman's rank, the p-value and correlation coefficient is identical for all three enzymes, and all correlations appear significantly strongly positively correlated with metabolite B (p-value < 0.05, therefore significantly different from 0).
Using Pearson's correlation test, the p-value and correlation coefficient is different, but all p-values are still far below the significance level (0.05, therefore significantly different from 0), so all appear to have a significant positive relationship when using this test as well.

B) Find the best-fit multiple-linear regression model for the dependence of metabolite on the enzyme concentrations and justify all your decisions with statistical evidence.

```
anova(lm(metaboliteB ~ enzymeA + enzymeB + enzymeC))
```

```
## Analysis of Variance Table
##
## Response: metaboliteB
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## enzymeA    1 7220876 7220876 514.8507 1.356e-13 ***
## enzymeB    1   44846   44846   3.1975    0.0927 .
## enzymeC    1    8336    8336   0.5944    0.4520
## Residuals 16  224403   14025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(metaboliteB ~ enzymeA + enzymeC + enzymeB))
```

```
## Analysis of Variance Table
##
## Response: metaboliteB
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## enzymeA    1 7220876 7220876 514.8507 1.356e-13 ***
## enzymeC    1    3001    3001   0.2139    0.6499
## enzymeB    1   50181   50181   3.5779    0.0768 .
## Residuals 16  224403   14025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(metaboliteB ~ enzymeC + enzymeA + enzymeB))
```

```
## Analysis of Variance Table
##
## Response: metaboliteB
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## enzymeC    1 6959475 6959475 496.2128 1.805e-13 ***
## enzymeA    1  264401  264401  18.8519 0.0005047 ***
## enzymeB    1   50181   50181   3.5779 0.0767964 .
## Residuals 16  224403   14025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(metaboliteB ~ enzymeC + enzymeB + enzymeA))
```

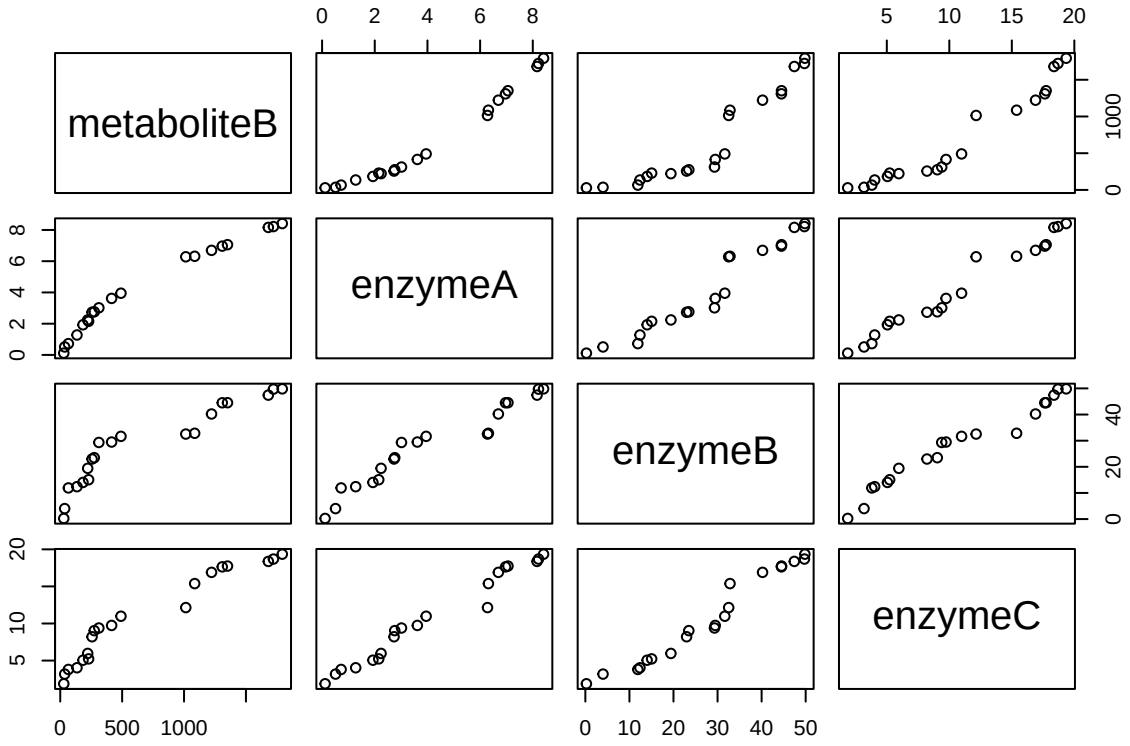
```
## Analysis of Variance Table
##
## Response: metaboliteB
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## enzymeC    1 6959475 6959475 496.2128 1.805e-13 ***
## enzymeB    1   34971   34971   2.4934 0.1338876
## enzymeA    1  279611  279611  19.9364 0.0003908 ***
## Residuals 16  224403   14025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As ANOVA tests in order, have repeated ANOVA with each combination of orders in the model. Significance determined by examining the F statistic (Pr(>F)) against the significance level of 0.05.

Enzyme B is not significant for any model (F stat is small), however enzyme C and enzyme A both appear significant for some arrangement of the model (where enzyme A is first in the model, enzyme C is not significant, but when enzyme C is first, enzyme C is significantly associated.) It appears that both enzyme A and enzyme C are dependent on metabolite B.

Next step: determine if enzyme C and A are independent (if yes, include both in final model, if no, only include the most significant variable)

```r
enzyme_table <- data.frame(metaboliteB, enzymeA, enzymeB, enzymeC)
pairs(enzyme_table)
```



```r
cor(enzyme_table, method="pearson")
```

```
##             metaboliteB   enzymeA   enzymeB   enzymeC
## metaboliteB   1.0000000 0.9813160 0.9292840 0.9633902
## enzymeA       0.9813160 1.0000000 0.9670427 0.9852242
## enzymeB       0.9292840 0.9670427 1.0000000 0.9790363
## enzymeC       0.9633902 0.9852242 0.9790363 1.0000000
```

From the Pearson's correlation test and the plots, enzyme A and enzyme C are significantly correlated with each other (not independent). Although enzyme B has already been discounted, enzyme B also looks dependent on (highly correlated against) both enzyme A and enzyme C. Therefore, only include one of these variables in the final model.

```r
reg_enzymeA <- lm(metaboliteB ~ enzymeA)
summary(reg_enzymeA) # r squared = 0.963
```

```
##
## Call:
## lm(formula = metaboliteB ~ enzymeA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158.35 -113.63  -21.61  115.53  222.49
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -218.86      50.40  -4.342 0.000393 ***
## enzymeA       219.03      10.12  21.639 2.46e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124.2 on 18 degrees of freedom
## Multiple R-squared:  0.963,  Adjusted R-squared:  0.9609
## F-statistic: 468.2 on 1 and 18 DF,  p-value: 2.463e-14
```

```r
reg_enzymeC <- lm(metaboliteB ~ enzymeC)
summary(reg_enzymeC) # r squared = 0.9281
```

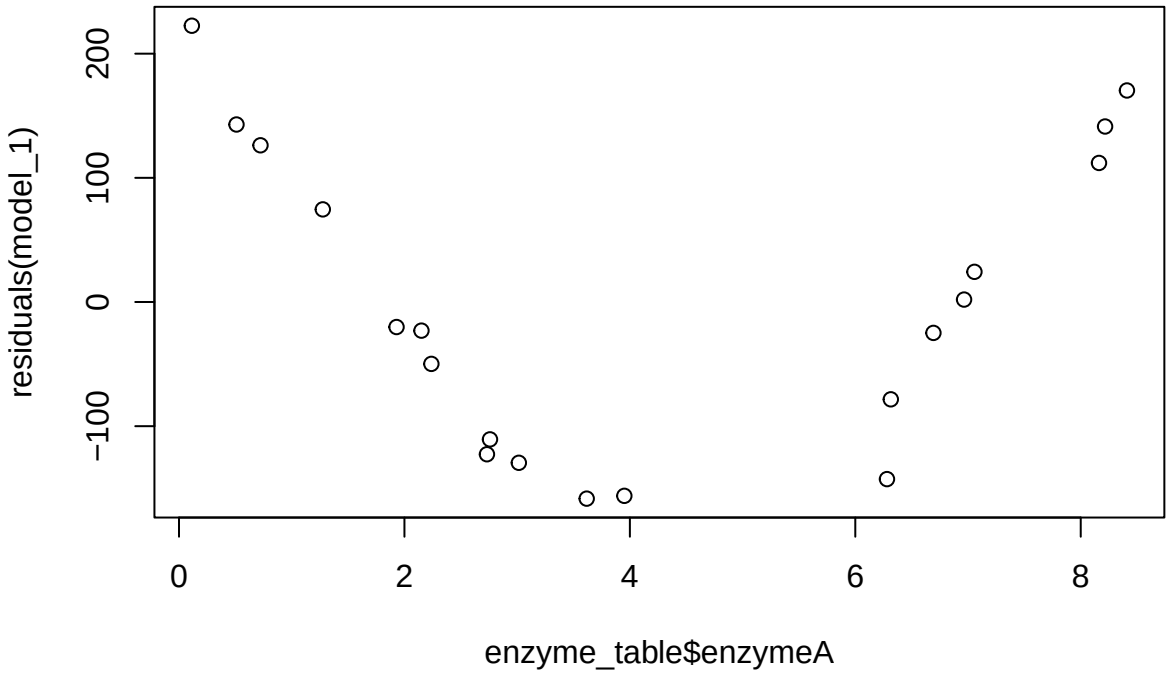
```
##
## Call:
## lm(formula = metaboliteB ~ enzymeC)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -255.0 -123.3   30.1  126.6  223.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -384.307     80.471  -4.776 0.000151 ***
## enzymeC      101.145      6.635  15.245 9.82e-12 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 173 on 18 degrees of freedom
## Multiple R-squared:  0.9281, Adjusted R-squared:  0.9241
## F-statistic: 232.4 on 1 and 18 DF,  p-value: 9.823e-12
```

Modelling each variable individually against metabolite B, both F statistics are significant (F statistics are large and p-values are below 0.05, therefore both variables are significant), but the correlation is stronger for enzyme A. Therefore, as enzyme A and enzyme C are dependent, the final model includes only enzyme A as a variable.

Next: is the model with just enzyme A linear? And if not, can this be improved?

```
model_1 <- reg_enzymeA

# residuals
plot(enzyme_table$enzymeA,residuals(model_1))
```




```
# adding a quadratic term
enzymeA_sq <- enzymeA*enzymeA
model_2 <- lm(metaboliteB ~ enzymeA + enzymeA_sq)
summary(model_1)
```

```
##
## Call:
```

```
## lm(formula = metaboliteB ~ enzymeA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158.35 -113.63  -21.61  115.53  222.49
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -218.86      50.40  -4.342 0.000393 ***
## enzymeA       219.03      10.12  21.639 2.46e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124.2 on 18 degrees of freedom
## Multiple R-squared:  0.963,  Adjusted R-squared:  0.9609
## F-statistic: 468.2 on 1 and 18 DF,  p-value: 2.463e-14
```

```
summary(model_2)
```

```
##
## Call:
## lm(formula = metaboliteB ~ enzymeA + enzymeA_sq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.602 -12.229  -1.759  13.520  37.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.984     14.801   2.026 0.058776 .
## enzymeA       34.527      8.600   4.015 0.000898 ***
## enzymeA_sq    20.886      0.949  22.008 6.23e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.53 on 17 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic:  6763 on 2 and 17 DF,  p-value: < 2.2e-16
```

```
anova(model_1,model_2)
```

```
## Analysis of Variance Table
##
## Model 1: metaboliteB ~ enzymeA
## Model 2: metaboliteB ~ enzymeA + enzymeA_sq
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     18 277585
## 2     17   9413  1    268172 484.34 6.234e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# adding a cubic term
enzymeA_cub <- enzymeA*enzymeA*enzymeA
model_3 <- lm(metaboliteB ~ enzymeA + enzymeA_sq + enzymeA_cub)
anova(model_2,model_3)
```

```
## Analysis of Variance Table
##
## Model 1: metaboliteB ~ enzymeA + enzymeA_sq
## Model 2: metaboliteB ~ enzymeA + enzymeA_sq + enzymeA_cub
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     17 9412.7
## 2     16 9403.6  1     9.165 0.0156 0.9022
```

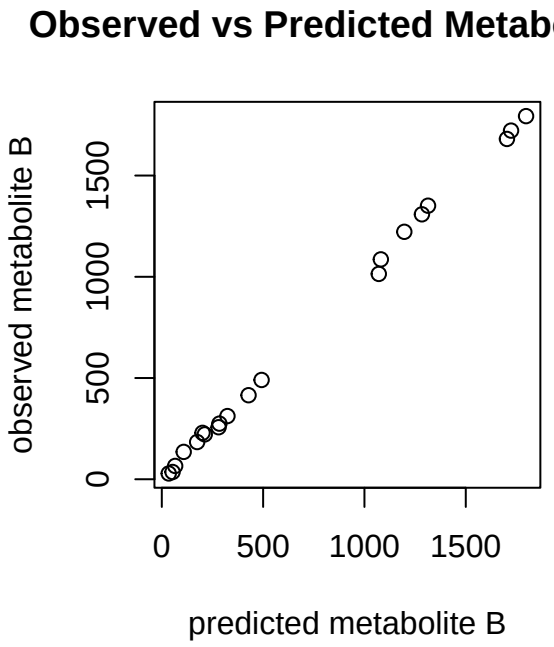
From the plot of residuals for the linear model (metabolite B and enzyme A), the functional form appears quadratic. Therefore, a quadratic term for enzyme A was added to a second model, which had a much larger F statistic which was also significant (F statistic for linear model: 468, for quadratic: 6763, p-value: <0.05), and a higher adjusted R squared (better fit). The partial F test in the ANOVA between these two models also showed significant improvement (<0.05).

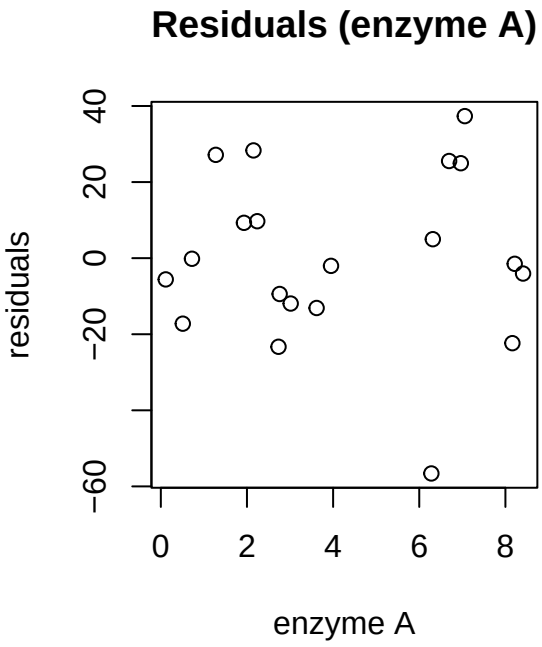
Adding a cubic term does not improve the model (significance for the partial F test is >0.05), therefore the best fitting model is model 2 - quadratic.

BEST FITTING MODEL FOR THE DATA: y = 29.984 + 34.527*enzymeA + 20.886*enzymeA*enzymeA

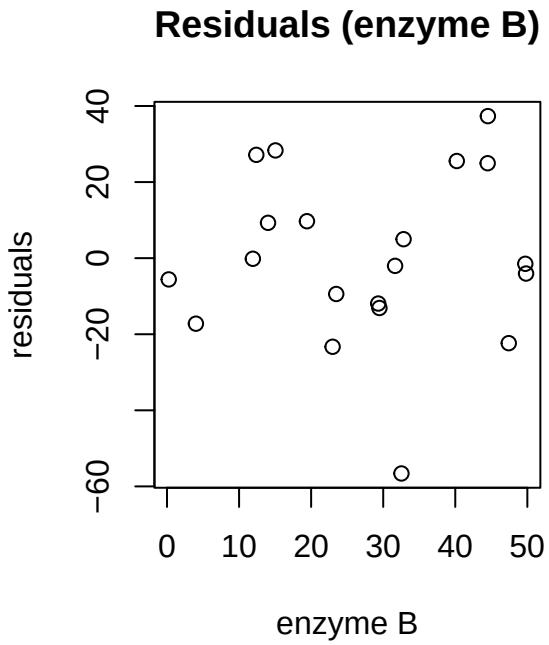
C) Create plots that demonstrate that your model is a good fit to the data

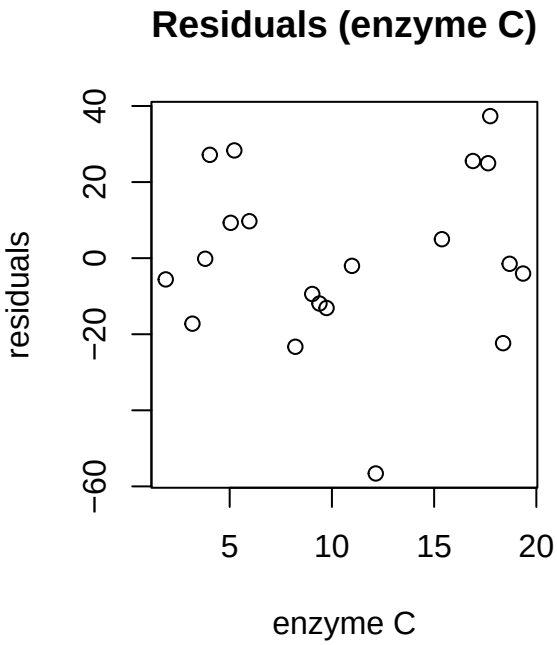
```r
par(mfrow=c(1,2), pty="s")
plot(29.984 + 34.527*enzymeA + 20.886*enzymeA*enzymeA, enzyme_table$metaboliteB,
     main = "Observed vs Predicted Metabolite B",
     ylab = "observed metabolite B",
     xlab = "predicted metabolite B")

# residuals
plot(enzyme_table$enzymeA,residuals(model_2),
     main = "Residuals (enzyme A)",
     ylab = "residuals",
     xlab = "enzyme A")
```






```
plot(enzyme_table$enzymeB,residuals(model_2),
     main = "Residuals (enzyme B)",
     ylab = "residuals",
     xlab = "enzyme B")
plot(enzyme_table$enzymeC,residuals(model_2),
     main = "Residuals (enzyme C)",
     ylab = "residuals",
     xlab = "enzyme C")
```


Residuals (enzyme B)
residuals
enzyme B
−60
−20
0
20
40
0
10
20
30
40
50


Residuals (enzyme C)
residuals
enzyme C
−60
−20
0
20
40
5
10
15
20